DOCUMENT RESUME

ED 410 284 TM 027 107

AUTHOR Reckase, Mark D.
TITLE Consequential Validity from the Test Developer's Perspective.
PUB DATE Mar 97
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *High School Students; High Schools; Higher Education; Ideology; Prediction; *Program Evaluation; Public Policy; *Test Construction; Test Use; Test Validity
IDENTIFIERS *ACT Assessment; Constructs

ABSTRACT

This paper explores what a responsible test developer would do to support the consequential validity of a test early in the development process, and how the consequential validity of the program should be monitored and addressed during the life of the program. To illustrate the issues concretely, the validity of the American College Testing Program (ACT) college entrance examination is considered as if the year were 1959 and the program were newly developed. Consequential validity is seen as having two dimensions: the appraisal of the value implications of the construct label, the theory underlying test interpretation, and the ideologies in which the theory is embedded; and the appraisal of the potential and actual social uses of the test. The ACT Assessment Battery was designed in the belief that the best predictor of future performance is a measure of past performance on tasks that are similar to the performance to be predicted. The ACT Assessment as designed appears to have construct labels, as test titles indicate, that meet the requirements of consequential validity in that they represent the test appropriately. The theory behind the test seems consistent with its uses. It is more difficult to evaluate the ideology behind the test (that a college education is an important goal and that students should prepare for it), but it is at least a recognizable ideology. Appraisal of the consequences of the test is more problematic, and it does not seem possible to meet the requirements of consequential validity because of the complications of the social policy implications of testing and unforeseen consequences of testing. (Contains one table and seven references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Mark Reckase
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

# Consequential Validity from the Test Developer's Perspective[1]

Mark D. Reckase
ACT, Inc.

ED 410 284

The consequential basis for validity as defined by Messick (1989) has been receiving increased attention in the educational measurement literature over the past few years (e.g., Messick, 1994; Messick, 1995; Moss, 1992), but at least from this author's perspective, little has appeared in the educational measurement literature that actually reports on all aspects of the consequential basis of validity for an existing testing program. Further, it is also unclear how the developer of a new testing program should collect information to support the consequential basis for validity of a testing program as it is being developed.

The purpose of this paper is to try to imagine what a responsible test developer would do to support the consequential validity of a test early in the development process, and how the consequential validity of the program should be monitored and addressed during the life of the program. There is special concern about the social aspects of consequential validity that require that the unanticipated consequences of a testing program be considered as part of the validation process.

To provide a concrete example of these issues, the validity of the ACT Assessment college entrance examination program will be considered as if it were 1959 and the program was newly developed. Since hindsight is much easier to consider than foresight, this example will

[1]Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, March, 1997.



TM027107

BEST COPY AVAILABLE

allow us to consider unanticipated consequences that are quite evident now, but that were not dreamed of in 1959.

Before considering the details of the issues, the definition of consequential validity that is used as the basis of this discussion is provided. While the term appears quite frequently in the literature, it is important for the discussion that the specifics of this relatively abstract concept be made clear.

## Definition of Consequential Validity

Messick (1989) indicated that the consequential basis of validity has two component parts and that they are distinct from the evidentiary basis for validity. He provides the following table (p.20) to help make the distinctions clear. The first component of consequential validity, the consequential basis of test interpretation "is the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and the ideologies in which the theory is embedded" (p.20). The second component, the consequential basis of test use, "is the appraisal of both potential and actual social consequences of applied testing" (p.20).

---

Insert Table 1 about here

---

These two components provide a very demanding set of criteria to be considered when attempting to provide support for the uses of a testing program. The first consideration under the consequential basis of test interpretation is that the labels given to test scores be evaluated to determine whether they "capture as closely as possible the essence of the construct's theoretical import (especially its empirically grounded import) in terms reflective of its salient value implications" (p.60). To be honest, I am not quite sure what that statement means in practice, but the examples in Messick's chapter suggest that test score labels should be accurate descriptions of the skills and knowledge assessed by a test and that they should not use language that has value loadings. For example, calling a mathematics achievement test the "World Class Mathematics Test" because it was reviewed by a few international scholars might be considered misleading and violating this component of consequential validity.

The second consideration under the first component is the theory underlying the test interpretation. The consequential basis for test interpretation requires the appraisal of the value implications of the theory. The third consideration is the ideologies in which the theory is embedded. It is difficult to imagine how the value implications of a theory or an ideology can be appraised. Certainly, there is the possibility that someone will develop a theory that persons with scores below a certain level are incapable of making moral judgements and should be institutionalized. This may be an unpalatable theory, but as with all testing enterprises, it should be supported by the evidentiary basis for validity. In other words, the values of the evaluator are no better than the values of the test constructor. Despite the difficulties in conceptualizing the

consequential basis for validity, an attempt will be made to appraise these features of consequential validity through the examples given below.

The second component of consequential validity, the appraisal of both potential and actual consequences of applying testing, also contains problems for the test developer. Certainly, when a test is in development, there are no actual consequences to appraise, there are only potential consequences. Some of those potential consequences can be anticipated from experience with other testing programs and the planned uses for the testing program. But, as I will show later, it is impossible to anticipate all consequences.

There is also a logical problem in identifying either type of consequence. The definition of a consequence is "the effect, result, or outcome of something occurring earlier" (Flexner, 1987). This definition implies that there is a cause and effect relationship between something that occurred earlier and the result. It is usually very difficult to demonstrate a cause and effect relationship unless there are carefully controlled experimental conditions. These typically are not present in a testing program. How are we to determine that any result is caused by the implementation of a testing program?

Based on this summary of the requirements for consequential validity, an attempt will be made to appraise the consequential validity of the ACT Assessment Program as if it were a new program in its early development.

## The ACT Assessment Test Battery

The ACT Assessment college admissions test battery was first administered in 1959. It was developed as a source of information that could be used by the advising staffs of the land grant colleges in the midwest to help place entering students in entry level courses that matched their educational background. At the time that the test battery was first used, the institutions that were using the test were not very selective in their admissions processes, so admissions was not a major use of the test.

The philosophy behind the design of the test is that the best predictor of future performance is a measure of past performance on a set of tasks that is similar to the performance to be predicted. That is, the best predictor of performance in college level courses is a measure of achievement in similar courses at the high school level. The tests were designed as a sample of tasks from the intersection of two domains. The first domain is the set of tasks that represent the knowledge and skills that are taught in grades nine through twelve. The second domain is the set of tasks that represent the knowledge and skills that are prerequisite to success in entry level college courses. The test construction process seeks to sample a set of tasks from the intersection of the two domains that provides good representation of subset defined by the intersection. The philosophy and details of the construction process of the early versions of the test are summarized in the program technical report titled "Assessing Students on the Way to College" (ACT, 1973). The testing program was extensively modified in 1989 so the descriptions in that document no longer apply. The revised version of the ACT Assessment Program is

described in the Preliminary Technical Manual for the Enhanced ACT Assessment (ACT, 1989). A new version of the technical manual for the program is currently in preparation.

The early version of the ACT Assessment Battery included four tests: English Usage, Mathematics Usage, Social Studies Reading, and Natural Sciences Reading. Each of these tests was designed to measure the level of acquisition of skills and knowledge taught in high schools during grades nine through twelve that were important for success in entry level college courses. ACT (1973) provides a list of expected uses for the program. They include: student self evaluation, college and general educational planning at the high school level, selection for admissions at the college level, course placement, and educational planning at the college level. At that time, little use was made of the test for scholarship selection. Such uses have been added to the expected uses since the early 1970s.

## The Consequential Validity of the Early ACT Assessment

### Construct Labels

Given the test design and philosophy, and the expected uses of the instrument, what can be said about its consequential validity using hindsight? First, the issue of construct labels will be considered. The overall title "ACT Assessment" seems pretty straight forward. It does not mislead, nor does it promise more than can be delivered.

The test names (construct labels) are a little more interesting. Many hours have been spent by ACT staff discussing whether a multiple-choice test could reasonably be called the English Usage Test. One perspective is that generation of English text is required before the test can be called "English Usage." Another perspective is that English is being used on the multiple-choice test, so the name is appropriate and accurate. Some have suggested a longer title -- "Some of the things that good writers need to know."

Similar discussions have addressed the Mathematics Usage Test. It does not assess all of mathematics usage, but only a sample of the possible topics. Perhaps it should be called "A sample of problems from the domain of mathematics that is prerequisite to success in entry level college courses." Of course, the more precise titles are unwieldy. The current version of the ACT Assessment labels these two tests simple the English Test and the Mathematics Test, but detailed documents are available for those that want to know exactly the skills and knowledge that are assessed by these tests. It would seem that these "construct labels" meet the requirements of consequential validity, but maybe someone believes that they imply something that the tests do not deliver.

**Theory**

The theory behind the ACT Assessment tests is domain sampling of a fairly well defined domain of content. The domain is identified by surveying secondary educators and determining what is taught at grades nine through twelve. This information is reviewed by college faculty

that teach entry level courses to identify the skills and knowledge that are assumed when students enter their classes. This process does have value implications. The high school curriculum is valued, but the critical component is the judgements of the college faculty. The theory seems consistent with the uses of the test battery.

**Ideology**

The test is imbedded within an ideology that suggests that a college education is an important goal and that students should prepare themselves to achieve that goal. There is also an implication that, for most students, certain fields of study are more direct prerequisites for success than other fields of study.

It is unclear how the theory behind the test and the ideological basis should be evaluated. Certainly there are other approaches to the development of college admissions tests. The SAT has a different theoretical basis. The means for appraising the value implications of the different approaches is unclear, except to make the theory and ideology clear so that users are aware of them.

**Consequences**

The appraisal of the potential and actual consequences of a test is somewhat more problematic. First there is the issue of whether the test causes some observed event. As part of the evidentiary part of validity, information is collected to support the use of the tests for placing

students in entry level courses. It is hoped that the students are more successful because of good placements. However, studies are not typically conducted that randomly assigns some students to classes while using test results to assign others students to the same classes. Thus, there is no causal evidence for the expected consequences, even if there is empirical evidence that the test does serve the desired purpose.

At a more global level, critics of testing practices have suggested that multiple choice tests, like the ACT Assessment "too often undermines vital social policies" (National Commission on Testing and Public Policy, 1990). Given the pervasive use of tests, it seems impossible to conclude that a cause and effect relationship exists between the use of a particular type of test and some social policy issue. How is the responsible testing organization to provide evidence that meets the requirements suggested by the consequential validity definition?

The issue of unanticipated consequences is even more difficult to address. When the ACT Assessment was first introduced, there were no professional coaching companies, and the NCAA was not using the test as part of the criteria for determining whether athletes could play on teams during their first year. Are those consequences caused by the introduction of the ACT Assessment? What should the test developer have done to appraise these consequences in 1959, or even 1973? It seems that it is impossible to meet the requirements set by the suggested consequential basis for validity, no matter how laudable meeting the requirements might be.

## An Afterword

While thinking about these issues, it occurred to meet that the consequential bases for validity apply to the definition of consequential validity as well as they do to testing issues. For example, what is the value implication of the label "consequential validity" and what are the theoretical and ideological underpinnings? How should those be appraised? How do we appraise the potential and actual consequences of using this concept of validity? I don't know the answers to these questions. I hope that they will be answered as part of this symposium.

## References


ACT (1973). *Assessing students on the way to college.* Iowa City, IA: The American College Testing Program.

ACT (1989). *Preliminary technical manual for the enhanced ACT Assessment.* Iowa City, IA: The American College Testing Program.

Flexner, S. B. (1987). *The random house dictionary of the English language* (2nd ed.). Unabridged. New York: Random House.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education, pp. 13-103.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

**Table 1**
**Facets of Validity**

| | Test Interpretation | Test Use |
|---|---|---|
| Evidential Basis | Construct validity | Construct validity + Relevance/utility |
| Consequential Basis | Value implications | Social consequences |

Messick, 1989